DOI:
10.1177/1742271X17753738

Document Version
Peer reviewed version

A comparison of ultrasound with magnetic resonance imaging in the assessment of fetal biometry and weight in the second trimester of pregnancy: An observer agreement and variability study.

Keywords

'Biometry', 'Fetal weight', 'Fetus', 'Observer variation', 'Magnetic Resonance Imaging', 'Ultrasonography', 'Pregnancy trimester, second'.

Abstract

Objective

To compare the intra and interobserver variability of ultrasound (US) and magnetic resonance imaging (MRI) in the assessment of common fetal biometry and estimated fetal weight (EFW) in the second trimester.

Methods

Retrospective measurements on pre-selected image planes were performed independently by two pairs of observers for contemporaneous US and MRI studies of the same fetus. Four common fetal measurements (BPD, HC, AC, FL) and an estimated fetal weight (EFW) were analysed for 44 'low risk' cases. Comparisons included, intra class correlation coefficients (ICC), systematic error in the mean differences, and the random error.

Results

The US inter- and intraobserver agreement were good, except for intraobserver AC (ICC = 0.880), with a significant increase in error with larger AC sizes. MRI produced excellent intraobserver agreement with higher ICCs than US. Good MRI interobserver agreement was comparable with US except for the BPD (ICC = 0.942, moderate). Systematic errors between modalities were seen for the BPD, FL and EFW (percentage error = +2.5%, -5.4% and -5.5% respectively, $p<0.05$). MRI had less random error than US for intraobserver HC, FL and EFW measures ($p<0.05$), and more interobserver error for the FL and EFW ($p<0.05$).

Conclusions

US remains the modality of choice when estimating fetal weight, however with increasing application of fetal MRI a method of assessing fetal weight is desirable. Both methods are subject to random error and operator dependence. Assessment of calliper placement variations, may be an objective method detecting larger than expected errors in fetal measurements.

Introduction

Accurate evaluation of fetal size and growth is essential for the delivery of good quality antenatal care, and ultrasound (US) measurements play a central role. When a US scan indicates that a fetus is appropriately grown this suggests good intrauterine health, thus is reassuring to the clinician and the parents. Additionally, an accurate antenatal detection of a growth abnormality may raise suspicions of a variety of fetal and maternal conditions which include; pre-eclampsia; fetal growth restriction, (FGR); gestational diabetes; macrosomia; infection; and syndromic or genetic conditions which are associated with changes in growth patterns (1,2). The information about fetal size is a marker of overall fetal health, and may act as a threshold for clinicians who, based on the findings, could offer further investigations such as Doppler US, blood tests, amniocentesis, or be used to plan the timing of delivery of a compromised fetus (3). However, ultrasound is known for its large random errors in fetal measurement and low sensitivity for detecting growth disturbances (2,4). Furthermore, there is growing evidence that magnetic resonance imaging (MRI) can result in estimated fetal weight (EFW) with far less error than ultrasound, particularly when using volumetric methods (5-7). Few studies have assessed the validity of MRI by radiologists for the measurement of fetal biometry compared to US by sonographers (8-10). Furthermore, a literature search revealed no studies which had performed a comprehensive intra- and inter-rater agreement, variability and method comparison of US and MRI for fetal biometry and estimated fetal weight (EFW). Additionally, reporting standards of method comparison studies vary widely which limits their interpretation (11-14).

Fetal MRI is a highly specialised modality for fetal diagnosis and is well established for fetal central nervous system (CNS) anomalies. A systematic review of 13 peer reviewed articles, found that MRI provided supplementary information to US and resulted in a change in clinical management in 30% of cases, with referral indications being numerous but including; posterior fossa anomalies, corpus callosal anomalies, microcephaly or apparently isolated ventriculomegaly (15,16). However, MRI is also increasing in its remit for fetal evaluation of anomalies outside the CNS e.g. diaphragmatic hernia, pulmonary anomalies and twin to twin transfusion syndrome, particularly when US is limited by reduced amniotic fluid, maternal obesity or in the presence of equivocal US findings (16-19). A survey conducted by the International

Society of Ultrasound in Obstetrics and Gynaecology (ISUOG), found that at least 1-2 centres in 27 countries were performing fetal MRI with the quality of imaging, sequences used and operator experience varying widely. In the UK, fetal MRI is offerd by some tertiary units who have a fetal medicine department (currently approximately 6 UK wide), and may involve outsourcing of image reporting to experienced specialists. ISUOG also suggests that a standardised and complete assessment of fetal anatomy is feasible with MRI, however its current remit is to complement an expert US examination (16).

As the use of clinical fetal MRI increases, modality specific biometric evaluation is becoming more important. Previous studies have almost exclusively focussed on fetal MRI late in gestation, however women may be referred for a fetal MRI scan soon after the 20 week anomaly US scan when anomalies are initially suspected (3,20). The aim of this study is to compare the intra and interobserver variability of US and MRI in the assessment of common fetal biometry and EFW in the second trimester. .

Design and Methods

The intelligent fetal imaging and diagnosis project (iFIND) is a large scale, single centre observational imaging and engineering project, whose aim is to use novel technologies to improve diagnosis and detection rates in the second trimester of pregnancy. The project has been granted NHS Research and Development approval and ethics approval, NRES reference number = 14/LO/1806, (trial registry numbers: UKCRN ID = 18283, ISRCTN = 16542843). All participants gave written and informed consent.

The study is divided into iFIND-1 where 10, 000 clinical mid-trimester anomaly ultrasound scans are recorded for the purposes of machine learning and big data analysis, and iFIND-2 which involves further imaging on a smaller population in addition to the anomaly scan. iFIND-2 includes a 2D and 3D US, as well as a MRI on each fetus, and these paired datasets are obtained within 0 to 3 days. The images were retrospectively and consecutively collected from the iFIND-2 datasets of fetuses with a normal anomaly scan result. The image planes pre-selected included the biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC) and femur length (FL) (see figures 1-8 for example image planes and measurement criteria). To calculate the EFW for each fetus (MRI and US) the Hadlock formula

including the HC, AC and FL measurements were used as recommended by the British Medical Ultrasound Society and ISUOG (20-22). Whilst the BPD measurement is useful to assess head shape its variability in measurement suggests it should not be used in routine EFW calculation (23).

The ultrasound system was a Philips Epiq (Philips Healthcare, Best, Netherlands) and the participants were examined by one of two operators (JM or CK), a Consortium for the Accreditation of Sonographic Education, CASE, accredited sonographer with 10 years scanning experience and a UK trained fetal medicine specialist with 6 years' experience respectively. A 6-1 mHz matrix probe was used to scan all patients. The MRI scanner used for all participants was a Philips Ingenia 1.5 Tesla system (Philips Healthcare, Best, Netherlands). Motion corrected MRI slice to volume reconstructions of the fetal head were used to find a transventricular plane comparable to US imaging (24).. An US and a MRI database of anonimysed paired scans was compiled using the Osirix image review software for off-line/remote review (version 7.5, Geneva, Switzerland). The databases were duplicated then the images reordered randomly, ready for a repeat review by the observers after 2.5 weeks, with the aim of reducing any recall bias. All reviewers were provided with face to face training and guidance notes about; which views to be recorded; the use of the Osirix review platform; and optimal viewing conditions for the off-line review.

Using both of the US databases, one sonographer (TF, a UK trained sonographer with 3 years scanning experience) performed repeated measures (blinded to MRI and any previous measurements), this was used for US intraobserver calculations. The fetal medicine specialist (CK) independently performed one US reading from the first database, for interobserver calculations. Using both the MRI databases one radiologist (KP, 5 years fetal MRI clinical experience) performed repeated measures (blinded to the US and any previous measures) and a fetal imaging research radiographer (CM, 10 years fetal MRI research experience) independently performed one MRI reading from the first database. The observers also recorded a 3 scale image quality score for each image (1=poor, 2=satisfactory and 3=good). Data was collected on an Excel spreadsheet and all supplementary materials and raw data was to be deposited at the University via a Research Data Management System on completion of analysis.

Figures 1 - 8: Image planes and measurement criteria

Image plane selection and calliper placement criteria was obtained from the NHS fetal anomaly screen programme guideline (20).

Head Circumference (HC), transventricular view

Figure 1: US HC plane

[place figure 1 here]

Figure 2: MRI SVR HC plane

[place figure 2 here]

In the transventricular view, the image plane was at the level of the cavum septum pellucidum anteriorly (*) and the lateral ventricular horn posterior containing the choroid plexus (^). The falx cerebri was mid-line (“) and the head an oval shape. The ellipse tool was used to measure around the outer table of the skull, being careful not to include any subcutaneous fat. The MRI transventricular view was carefully selected from slice to volume reconstructions (SVR) (24) obtained from T2 dynamic sequences (TR/TE = Longest/80, slice Th/gap = Volume/-1.25) which were manipulated in Osirix using the multiplanar reconstruction (MPR) mode.

Biparietal Diameter (BPD), transventricular view

Figure 3: US BPD plane

[place figure 3 here]

Figure 4: MRI SVR BPD plane

[place figure 4 here]

In the same image plane as the for the HC measurement, the BPD was measured from the outer table of the skull to the outer table of the skull at the widest part for both MRI and US.

Abdominal Circumference (AC)

Figure 5: US AC plane

[place figure 5 here]

Figure 6: MRI AC plane

[place figure 6 here]

The AC measurement was obtained with an ellipse tracing. The image plane was at a level including the part of the fetal liver (*), the fetal stomach (^), the portal sinus of the umbilical vein (“), 3 bony points of a vertebra in cross section (+), a circular abdominal appearance, circular aorta (>) and with a short length of a rib, i.e. ‘unbroken’ (‘). The MRI sequence most commonly selected with the correct plane, was a T2 fast spin echo sequence of the transverse uterus (TR/TE = 920/90, slice Th/gap = 4/0), followed by the single shot fast spin echo.

Femur Length (FL)

Figure 7: US FL plane

[place figure 7 here]

Figure 8: MRI FL plane

[place figure 8 here]

The FL was measured by placing the callipers at the end of the diaphysis in a view where the femur does not appear foreshortened (solid line). Care was taken to avoid measuring the cartilaginous epiphysis at either end of the femur and also to avoid the greater trochanter which otherwise would falsely elongate the measurement. The MRI sequence most commonly found to have a clear view of the femur in the correct plane was a DWI sequence in the B0 field i.e. before the diffusion weighting was applied, (TR/TE = 4000/89, slice Th/gap = 5/0). Some MRI femur views were well visualised using a gradient echo echoplanar imaging sequence.

Statistical analysis

The data was analysed using the statistical packages, SPSS (version 23, SPSS Inc, Chicago, Ill, USA) and Excel, (version 14.4.7, Microsoft Corp. Redmond, Washington, USA). The EFW was calculated using the Hadlock formula (25). A power calculation determined that a sample size of 31 was required to give a power of 80% for an error of 5% to detect an effect size of 1 mm difference (assuming a standard deviation of 8mm). Normality testing was performed to ensure assumptions were met for statistical analysis and to identify any obvious outliers.

To assess systematic error between the modalities, the mean difference in measurement from the two observers per modality was compared for each parameter (BPD, HC, AC, FL, and EFW). A two tailed paired t-test was performed to compare the means.

To test the intra and interobserver agreement, the average measures intra class correlation coefficient, ICC was used. Suggested cut off limits proposed in the literature for fetal studies guided interpretation (26).

Bland Altman plots were used to graphically assess the mean difference and the limits of agreement, LoA. A linear regression coefficient was used to determine if there was a statistically significant proportional bias in the error as the size increased.

Random error was compared between modalities using the LoA (+/-1.96SD of the mean) as a marker of intra and interobserver variability and a two tailed paired t-test was performed.

Finally, to allow the clinical significance to be interpreted more readily, the proportion of cases falling outside of a calliper placement error threshold was calculated. Arbitrary thresholds were determined by previous examples of expected error in the literature (4). In addition, a standard deviation (SD) threshold for each parameter was determined using 1SD of the US intraobserver measurements observed. A number and percentage of cases falling outside of the threshold ranges were tabulated and compared between MRI and US.

Results

53 consecutive iFIND-2 participants were recruited between November 2015 and April 2016 and had their fetal imaging studies reviewed for inclusion. 44 participants

(83%) had fully paired datasets, and of these 25 (47%) had complete datasets and 19 (36%) were partially complete. Nine cases were excluded from the study because: four did not attend both scans; two had no transventricular US scan plane available; two had failed or poor quality MRI head SVRs; and two had missing US images.

The gestational age, (GA), was a mean of 23.5 weeks (range 20.3 – 25.7). The BMI was a mean of 26.3kg/cm (range 22.2 – 38.4kg/cm), with 3 cases above 30kg/cm (clinically obese). 68% of US and MRI scans were on the same day, 4% had a 2 day interval and 24% had a 3 day interval. 84% of the US scans had a satisfactory mean image score and 16% had a good score. For MRI, 8% had a poor mean score, 80% had a satisfactory score and 12% had a good score.

Table 1: Difference in the mean US and MRI biometric measurements and EFW

Measurement	**n**	**US, Mean, mm**	**MRI_{biom}, Mean, mm**	**Absolute difference, mm (95% CI)**	**Percentage difference, %**	**Paired t-test (p-value)**
BPD	30	58.8	60.2	-1.5 (-2.2 - -0.8)	-2.5	<0.001
HC	30	215.5	215.4	0.6 (-1.4 - 1.5)	0.3	0.9
AC	42	191.4	190.3	1.1 (-2.3 – 4.5)	0.6	0.5
FL	33	42.0	39.7	2.2 (1.0 – 3.7)	5.4	0.001
EFW, g	25	647.1	593.3	53.8 (19-89)	8.7	<0.05

Table 1 demonstrates that MRI systematically measured the BPD larger than US (mean percentage error = 2.5%, or 1.5mm, p = 0.001), and the FL smaller than US (mean percentage error = -5.4%, or -2.2mm, p = 0.001). MRI systematically measured the EFW smaller than US, (mean percentage error = -5.5%, or -34.8g, $p<0.05$). The mean measurements of the HC and AC compared well between modalities.

After normality testing, two outliers were removed from the dataset for the subsequent analysis. One was an obvious data input error for the MRI BPD (case 6) and one was a significant measurement error due poor image quality of a T2 sequence for bone (case 18). Only one other outlier was identified for US AC, however it was

unclear if this was a data input error or a true observer measurement so was kept for the remaining analysis (case 41, see figure 13).

Table 2: Intraobserver and interobserver agreement, ICC

Fetal Measurement (n)	**US ICC (95% CI)**	**MRI ICC (95%CI)**
Intraobserver		
BPD (28)	0.982, good, (0.959 - 0.992)	0.995, excellent, (0.988 - 0.997)
HC (28)	0.952, good, (0.580 - 0.986)	0.997, excellent, (0.994 - 0.999)
AC (40)	0.880, poor, (0.772 -0.937) *Significant proportional bias, p <0.05*	0.994, excellent, (0.988 - 0.997)
FL (31)	0.978, good, (0.944- 0.990)	0.989, good, (0.975 -0.995)
EFW (23)	0.972, good, (0.547 - 0.993)	0.983, good, (0.961 – 0.993)
Interobserver		
BPD (28)	0.974, good, (0.808 - 0.992)	0.942, moderate, (0.860 – 0.974)
HC (28)	0.971, good, (0.938 - 0.987)	0.983, good, (0.963 -0.992)
AC (40)	0.967, good, (0.896 -0.982)	0.973, good, (0.949 -0.986)
FL (31)	0.990, good, (0.979 -0.995)	0.978, good, (0.955 -0.990)
EFW (23)	0.988, good, (0.965 - 0.995)	0.964, good, (0.905 -0.986)

Table 2 shows that MRI had excellent intraobserver agreement for BPD, HC, AC, EFW and good FL agreement, with all ICC results scoring higher than US. Only the intraobserver FL and EFW had overlapping confidence intervals between modalities suggesting significant differences in agreement for the remaining biometry. US had good intraobserver agreement for all parameters except AC which scored poorly (ICC = 0.880). In addition, there was significantly less agreement for the US AC intraobserver measurement as the AC absolute size increased ($p < 0.05$).

For interobserver agreement US and MRI both had good agreement for all parameters except for the MRI BPD (moderate ICC = 0.942), however all parameters had overlapping 95% confidence intervals, suggesting no significant difference.

Figure 9-18: Bland Altman Plots of US compared to MRI, showing mean absolute error, mm, and limits of agreement, LoA, (+/- 1.96 SD) above and below the mean.
US = blue circles, o and solid lines ——— , MRI = green crosses, x and dashed line --------

Figure 9 Intraobserver BPD

[place figure 9 here]

Figure 10 Interobserver BPD

[place figure 10 here]

Figure 11 Intraobserver HC

[place figure 11 here]

Figure 12 Interobserver HC

[place figure 12 here]

Figure 13 Intraobserver AC

[place figure 13 here]

Figure 14 Interobserver AC

[place figure 14 here]

Figure 15 Intraobserver FL

[place figure 15 here]

Figure 15 Interobserver FL

[place figure 16 here]

Figure 17 Intraobserver EFW

[place figure 17 here]

Figure 18 Interobserver EFW

[place figure 18 here]

The Bland Altman plots in Figures 9-18 shows the absolute difference in millimeters between two measurements for each individual case. The MRI and US differences are overlaid on the same plot with a central mean difference line and a limits of agreement line above and below to represent 95% of the variance. Only intraobserver AC showed an increase in variation with size, with a marginal increased seen with intraobserver FL that was not significant. The LoA varied between parameters, with a tendency for MRI LoA to be narrower than US for intraobserver measures and wider for interobserver measures.

In table 3, the LoA (random error) are explored further, and demonstrates that statistically significant differences were observed for the intraobserver LoA for HC, FL and EFW, with MRI having less variation than US ($P<0.05$). There were significant differences in the interobserver LoA for AC and FL, with MRI having

more variation than US ($p < 0.05$). Parameters where the mean variation was above an arbitrary 5% percentage error threshold, included the intraobserver US measures of AC, FL and EFW (8.7%, 5.0% and 6.6% respectively) and MRI EFW (6.2%). For interobserver measures, the parameters for MRI with a mean percentage error above 5% include BPD, AC, FL and EFW (5.0%, 5.5%, 6.9% and 10.1% respectively) . For US, only interobserver EFW had and a mean percentage error of more than 5% (6.2%).

Table 3: Differences in random error between US and MRI fetal measurements and biometry derived EFW (paired t-test)

Fetal measurement (n)	**Intraobserver**			**Interobserver**		
	US	**MRI**	**p-value**	**US**	**MRI**	**p-value**
Absolute error, mm, (+/- 1.96 SD)						
BPD (28)	1.4	1.1	0.3	1.3	3.1	0.8
HC (28)	7.1	2.5	<0.05	7.9	6.2	0.6
AC (40)	18.0	5.1	0.09	8.7	10.3	<0.05
FL (31)	2.1	1.6	<0.05	1.6	2.6	0.99
EFW, g (23)	45.6	73.2	0.1	43.0	54.2	0.6
Percentage error, %, (+/- 1.96 SD)						
BPD (28)	2.4%	1.8%	0.3	2.2%	5.0%	0.9
HC (28)	3.2%	1.2%	<0.05	3.6%	2.8%	0.6
AC (40)	8.7%	2.7%	0.1	4.6%	5.5%	<0.05
FL (31)	5.0%	4.2%	<0.05	3.8%	6.9%	0.97
EFW, g (23)	6.5%	13.6%	0.2	6.3%	8.9%	0.8

Table 4: Differences in proportion of US and MRI cases falling outside of arbitrary error threshold

Arbitary cut off			**Intraobserver > threshold**		**Interobserver > threshold**	
Parameter	**Total cases**	**Threshold values. Intra/inter (mm)**	**US = n (%)**	**MRI n (%)**	**US n (%)**	**MRI n (%)**
BPD	28	1.4/2.2	1 (4)	1 (4)	0 (0)	6 (21)
HC	28	5.2/8.0	11 (39)	0 (0)	2 (7)	1 (4)
AC	40	7.9/11.0	16 (15)	0 (0)	1 (3)	3 (8)
FL	31	2.1/2.5	3 (10)	0 (0)	0 (0)	2 (6)
EFW	23	66.1g (10%)	2 (13)	3 (0)	0 (0)	0 (0)
Total number of cases measures of range (n=150)			33 (22)	4 (3)	3 (2)	12 (8)

Table 5: Differences in proportion of US and MRI cases falling outside of 1 SD error threshold

1 SD cut off			Intraobserver > threshold		Interobserver > threshold	
Parameter	**Total cases**	**Threshold values. mm**	**US = n (%)**	**MRI n (%)**	**US n (%)**	**MRI n (%)**
BPD	28	0.6	13 (46)	8 (29)	16 (57)	17 (61)
HC	28	2.8	15 (54)	1 (4)	10 (36)	6 (21)
AC	40	3.3	19 (48)	8 (20)	21 (53)	17 (43)
FL	31	1.1	7 (23)	6 (19)	5 (16)	9 (29)
EFW	23	33.1g (5%)	8 (35)	10 (43)	4 (17)	6 (26)
Total number of measures out of range (n = 150)			62 (41)	33 (22)	56 (37)	55 (37)

Table 4 demonstrates that more US cases that fell outside of the anticipated error range when compared to MRI (32 US cases versus 1 case for MRI), with MRI performing equal to, or better than, US for all parameters. For interobserver error 15 MRI cases and 3 US cases fell outside the expected threshold for error, with US performing better than MRI for BPD, AC, FL and EFW.

Table 5, with narrower thresholds (based on intraobserver US SDs), demonstrated MRI measurements that consistently had less cases falling out of range compared to US for intraobserver measures (62 US cases versus 33 MRI cases). For interobserver cases there were 56 US cases and 55 MRI cases in total with larger error. For intraobserver EFW with SD thresholds, the MRI measurements appeared to perform better than US with less cases with large variations i.e. >5% (8 cases or 35%, versus, 2 cases or 9%).

Discussion

This study sought to comprehensively compare the intra- and interobserver variability between MRI and US for fetal measurements and EFW. The calliper placement error for both US and MRI were found to be small (less than 5%), however the random errors observed were expected to be smaller than in clinical practice because of the highly controlled conditions (one image plane selected per participant and low risk fetuses), thus should be interpreted with caution. US was more susceptible to intraobserver variability, whereas MRI was more susceptible to interobserver

variability, both having cases falling outside of previously published error thresholds for fetal measurements (4). The causes of random errors in the US measurements that are used to calculate EFW, are multifactorial in origin and include; fetal position; maternal adiposity; sonographer experience; equipment specification; and reduced amniotic fluid which could limit the view (4,27,28). Observer variation, is known to have a major impact on the precision of US fetal measurements, with electronic calliper placement on an image, accounting for 58-80% of the error, having more impact than maternal adiposity or fetal position (4,5). This highlights the need for thorough operator training and audit but also the need for technological development of more quantifiable and less subjective assessments (29).

Sarris et al in 2012, investigated fetal biometry variation in 175 cases with three experienced sonographer observers, and found intraobserver variation to be consistently smaller than intraobserver variation. The poorer US intraobserver measurements (compared to inter- variation) observed in this study was surprising, and highlights the need for objective measurement audits in departments on an individual basis because this has a direct impact clinically when serial scans are performed often by different operators. For MRI, the wider interobserver error was expected as these fetal measurements are rarely measured routinely and the operator experience thus limited. Fetal MRI staff not experienced in performing these measurements will need more training in the future and there is a case for objective validation and also for US and MRI specialists to work across disciplines, developing practice that compliment one another. MRI currently has no universally agreed modality specific growth charts validated for clinical use, largely because; MRI is a relatively new tool with less reference data available; most fetal MRI examinations are for the brain or spine where the technique is better established, and; there is an assumption that the routinely utilised US reference data and growth charts are suitable to use across the two modalities (9,30).

The larger US intraobserver variation and increased variation with increased size for the AC measurements, from which the EFW formulae are based, suggests a measurement that should be closely monitored. Previous studies have supported the finding that AC measurements have less variation than EFW and therefore be a better predictor of size at term (31), however here we demonstrate that the role of the

operator is still very important. A tool to assess calliper placementt, monitoring groups of ultrasound operators, could use z-scores or relative percentage error to assess departmental and individual variance across time or as a training tool (4). Whether using US or MRI, operator dependence in obtaining fetal biometry reaffirms the importance of quality training and audit to reduce random errors and recommendations have been published in the literature for sonographers (28,32-34).

The EFW variability suggests that the random errors in fetal measurements will often compound the systematic errors of the mathematical equation, whether using US or MRI (35). Indeed, Khel et al, 2012 suggests that the current accuracy of EFW has reached its limits, and that novel approaches to US technology must be considered to reduce clinical errors. 3D US volumes of a part of a fetusus' limb, which incorporates soft tissue, has been used in EFW calculations, with some success, to improve accuracy, however as yet, there is a paucity of diagnostic accuracy tests to validate its use clinically (27,36-38), and reductions in post processing time is needed to make this a useful tool in the future (1,2). Significant variation in EFW calculations has clinical implications because currently US is not recommended to screen the low risk population for growth disturbances due to poor sensitivity and specificity (39). Additionally, errors in the formula occurs at the extremes of the weight range, due to changes in the soft tissue fat/muscle ratio of a compromised fetus, and may result in an overestimation of weight in small babies and an underestimation of weight in large babies when accurate depiction is most clinically important (40).

There is growing evidence that volumetric magnetic resonance imaging (MRI) can result in EFWs compared to birthweight with less random error than US, reported as low as 1-3% versus up to 7% for US (5,9,41,42). MRI is well established as a multiplanar reconstruction (MPR) imaging technique which means that two-dimensional images from a 3D volume of data can be reconstructed and viewed in any orientation, thus theoretically reducing measurement errors caused by an oblique or off centre plane. Moreover, MRI can negate some of US's technical drawbacks because maternal size, amniotic fluid and fetal position are less of a problem due to MRIs increased field of view. Still, fetal movements in MRI can cause image degradation, particularly at earlier gestations when the fetus is more active, and results in a poorer signal-to-noise ratio. However, MRI has superior soft tissue

contrast and improved boundary definition when placing electronic callipers for measurement or when outlining segments of the fetal body to calculate a volume.

Although MRI is underused as an antenatal tool compared to US, it is increasing its remit within fetal imaging for complex or equivocal cases due to improvements in post processing and faster scan sequences that reduce the issues of fetal movement MRI, and it is based on non-ionising radiation and is considered safe to use in pregnancy (43,44). Nonetheless, the use of MRI is limited by its expense, lack of expertise and scanner availability, as well as the limited evidence base of MRI's advantages over obstetric US for non-central nervous system anomalies.Differences in the imaging physics of each method are likely to account for the systematic error in the mean measurement between modalities (9,11). For example, the use of T2 weighted MRI images could mean the anatomical landmarks are slightly different to US, e.g. more subcutaneous scalp tissue may have been included due to the poorer bone definition. Distortion effects of the echoplanar imaging sequences used to select a FL plane on MRI may have resulted in the smaller FL measured. Technical refinement of MRI sequences may be necessary for a comparable and representative assessment of fetal anatomy

A major strength of the study was the adherence to current literature on reliability and agreement studies, the use of recommended statistics and guidance on interpretation, thus avoiding some of the heterogenous methods used in previous publications (11-13,26). As a retrospective study, limitations in the sample size occured due to the availability applicable of datasets, the short timeframe for the study, and lack of control over image quality was an issue. Also, a prospective study would mean real time US (as in clinical practice) could reveal the true variability. Furthermore, US was used as the reference standard to compare MRI – however it is well documented that the technique is prone to errors Due to the small numbers no statistical assessment of confounders (e.g. BMI, or fetal position) could be attempted. Furthermore, it may have been helpful to report the findings in terms of gestational age to aid easy interpretation by the clinician.

Future research should investigate the role of whole fetal body volume segmentation by MRI (or US) in the assessment of fetal weight as the technology continues to develop at a rapid pace (5,27,36). Methods to assess measurement variability as part

of individual and departmental audit should also be investigated as part of audit or training programmes, with the aim of providing much needed objective quality assurance.

Conclusion

US remains the modality of choice when assessing biometry and estimating fetal weight. However with increasing applications of fetal MRI, a method of assessing fetal growth and weight is desirable. Both methods are subject to random error and operator dependence, with US being more operator dependant and MRI being an immature modality for common biometry. Since, EFW is affected by the variability of 2D measures, novel approaches, such as 3D volumetric methods in MRI, need further investigation if clinical errors are to be reduced in the future. The assessment of calliper placement variations, may be an objective method detecting larger than expected errors in fetal measurements.

References

(1) Malin G.L., Bugg G.J., Takwoingi Y., Thornton J.G., Jones NW. Antenatal magnetic resonance imaging versus ultrasound for predicting neonatal macrosomia: A systematic review and meta-analysis. BJOG: An International Journal of Obstetrics and Gynaecology 2016 01 Jan 2016;123(1):77-88.

(2) Dudley N. A review of ultrasound fetal weight estimation in the early prediction of low birthweight. Ultrasound 2013 November 01;21(4):181-186.

(3) RCOG. Termination of Pregnancy for Fetal Abnormality. A working party report. Royal College of Obstetricians and Gynaecologists 2010:Accessed from https://www.rcog.org.uk/globalassets/documents/guidelines/terminationpregnancyreport18may2010.pdf on 16/04/2016.

(4) Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, et al. Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet Gynecol 2012;39(3):266-273.

(5) Kacem Y, Cannie MM, Kadji C, Dobrescu O, Lo Zito L, Ziane S, et al. Fetal weight estimation: comparison of two-dimensional US and MR imaging assessments. Radiology 2013 Jun;267(3):902-910.

(6) Zaretsky M.V., Reichel T.F., McIntire D.D., Twickler DM. Comparison of magnetic resonance imaging to ultrasound in the estimation of birth weight at term. Obstet Gynecol 2003 October 2003;189(4):1017-1020.

(7) Baker P.N., Johnson I.R., Gowland P.A., Hykin J., Harvey P.R., Freeman A., et al. Fetal weight estimation by echo-planar magnetic resonance imaging. Lancet 1994 1994;343(8898):644-645.

(8) James JR, Khan MA, Joyner DA, Buciuc B, Bofill JA, Liechty KW. MR Biomarkers of Gestational Age in the Human Fetus. Magnatom Flash 2012;1:Accessed at: www.siemens.com/magnetom-world on 30/11/2015-p112-118.

(9) Parkar AP, Olsen OE, Gjelland K, Kiserud T, Rosendahl K. Common fetal measurements: a comparison between ultrasound and magnetic resonance imaging. Acta Radiol 2010 Feb;51(1):85-91.

(10) Hatab MR, Zaretsky MV, Alexander JM, Twickler DM. Comparison of fetal biometric values with sonographic and 3D reconstruction MRI in term gestations. AJR Am J Roentgenol 2008 Aug;191(2):340-345.

(11) Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound in Obstetrics and Gynecology 2008;31(4):466-475.

(12) Coelho Neto MA, Roncato P, Nastri CO, Martins WP. True Reproducibility of UltraSound Techniques (TRUST): Systematic review of reliability studies in obstetrics and gynecology. Ultrasound Obstet Gynecol 2015;46(1):14-20.

(13) Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound in Obstetrics and Gynecology 2003;22(1):85-93.

(14) Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310.

(15) Rossi AC, Prefumo F. Additional value of fetal magnetic resonance imaging in the prenatal diagnosis of central nervous system anomalies: a systematic review of the literature. Ultrasound Obstet Gynecol 2014 Oct;44(4):388-393.

(16) Prayer D, Malinger G, Brugger PC, Cassady C, De Catte L, De Keersmaecker B, et al. ISUOG Practice Guidelines: performance of fetal magnetic resonance imaging. Ultrasound in Obstetrics & Gynecology 2017;49(5):671-680.

(17) Gonçalves LF, Lee W, Mody S, Shetty A, Sangi-Haghpeykar H, Romero R. Diagnostic accuracy of ultrasonography and magnetic resonance imaging for the detection of fetal anomalies: a blinded case–control study. Ultrasound in Obstetrics & Gynecology 2016:n/a-n/a.

(18) Pugash D. Fetal MRI: the sonographer's view. Top Magn Reson Imaging 2011 Jun;22(3):91-99.

(19) Gholipour A, Estroff JA, Barnewolt CE, Robertson RL, Grant PE, Gagoski B, et al. Fetal MRI: A technical update with educational aspirations. Concepts Magn Reson Part A Bridging Educ Res 2014;43(6):237-266.

(20) PHE. Fetal Anomaly Screening Programme Standards 2015-16. Public Health England 2015:Accessed at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/421650/FASP_Standards_April_2015_final_2_.pdf on 27/04/16.

(21) Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with the use of head, body, and femur measurements--a prospective study. Am J Obstet Gynecol 1985 Feb 1;151(3):333-337.

(22) Salomon LJ, Alfirevic Z, Berghella V, Bilardo C, Hernandez-Andrade E, Johnsen SL, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics & Gynecology 2011;37(1):116-126.

(23) Loughna P, Chitty L, Evans T, Chudleigh T. Fetal size and dating: charts recommended for clinical obstetric practice. Ultrasound 2009;17(3):161-167.

(24) Keraudren K, Kuklisova-Murgasova M, Kyriakopoulou V, Malamateniou C, Rutherford MA, Kainz B, et al. Automated fetal brain segmentation from 2D MRI slices for motion correction. Neuroimage 2014 Nov 1;101:633-643.

(25) Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with the use of head, body, and femur measurements--a prospective study. Am J Obstet Gynecol 1985 Feb 1;151(3):333-337.

(26) Martins WP, Nastri CO. Interpreting reproducibility results for ultrasound measurements. Ultrasound in Obstetrics & Gynecology 2014;43(4):479-480.

(27) Kehl S, Schmidt U, Spaich S, Schild RL, Sutterlin M, Siemer J. What are the limits of accuracy in fetal weight estimation with conventional biometry in two-dimensional ultrasound? A novel postpartum study. Ultrasound Obstet Gynecol 2012 May;39(5):543-548.

(28) Dudley N, Russell S, Ward B, Hoskins P, BMUS QA Working Party. BMUS guidelines for the regular quality assurance testing of ultrasound scanners by sonographers. Ultrasound 2014 February 01;22(1):8-14.

(29) PHAST/DoH. Errors in Epidimological Measurement. Health Knowledge - Public Health Action Support Team 2011:Accessed at: http://www.healthknowledge.org.uk/e-learning/epidemiology/practitioners/errors-epidemiological-measurements on 16/06/2016.

(30) Kyriakopoulou V, Vatansever D, Davidson A, Patkee P, Elkommos S, Chew A, et al. Normative biometry of the fetal brain using magnetic resonance imaging. Brain Structure and Function 2017 07/01;222(5):2295-2307.

(31) Nesbittâ€Hawes EM, Tetstall E, Gee K, Welsh AW. Ultrasound (in)accuracy: it's in the formulae not in the technique â€“ assessment of accuracy of abdominal circumference measurement in term pregnancies. Australas J Ultrasound Med 2014 Feb;17(1):38-44.

(32) Dudley NJ, Chapman E. The importance of quality management in fetal measurement. Ultrasound Obstet Gynecol 2002 Feb;19(2):190-196.

(33) Royal College of Radiologists, the Society and College of Radiographers. Standards for the provision of an ultrasound service. London: The Royal College of Radiologists; 2014.

(34) Society and College of Radiographers, British Medical Ultrasound Society. Guidelines for professional ultrasound practice. London: SCoR/BMUS; 2016 (rev. 1).

(35) Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. Ultrasound in Obstetrics and Gynecology 2005;25(1):80-89.

(36) Schild RL. Three-dimensional volumetry and fetal weight measurement. Ultrasound in Obstetrics and Gynecology 2007;30(6):799-803.

(37) Bennini JR, Marussi EF, Barini R, Faro C, Peralta CF. Birth-weight prediction by two- and three-dimensional ultrasound imaging. Ultrasound Obstet Gynecol 2010 Apr;35(4):426-433.

(38) Lima JC, Miyague AH, Filho FM, Nastri CO, Martins WP. Biometry and fetal weight estimation by two-dimensional and three-dimensional ultrasonography: an intraobserver and interobserver reliability and agreement study. Ultrasound in Obstetrics & Gynecology 2012;40(2):186-193.

(39) NICE. Antenatal Care for Uncomplicated Pregnancies. National Institute for Clinical Excellence 2008:Accessed at: https://www.nice.org.uk/guidance/cg62/chapter/1-guidance#fetal-growth-and-wellbeing on 27/04/2016.

(40) Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. Ultrasound Obstet Gynecol 2005 Jan;25(1):80-89.

(41) Hatab MR, Zaretsky MV, Alexander JM, Twickler DM. Comparison of fetal biometric values with sonographic and 3D reconstruction MRI in term gestations. Am J Roentgenol 2008;191(2):340-345.

(42) Uotila J., Dastidar P., Heinonen T., Ryymin P., Punnonen R., Laasonen E. Magnetic resonance imaging compared to ultrasonography in fetal weight and volume estimation in diabetic and normal pregnancy. Acta Obstet Gynecol Scand 2000 2000;79(4):255-259.

(43) Reddy UM, Abuhamad AZ, Levine D, Saade GR, Fetal Imaging Workshop Invited Participants. Fetal imaging: executive summary of a joint Eunice Kennedy Shriver National Institute of Child Health and Human Development, Society for Maternal-Fetal Medicine, American Institute of Ultrasound in Medicine, American College of Obstetricians and Gynecologists, American College of Radiology, Society for Pediatric Radiology, and Society of Radiologists in Ultrasound Fetal Imaging workshop. Obstet Gynecol 2014 May;123(5):1070-1082.

(44) Plunk MR, Chapman T. The Fundamentals of Fetal MR Imaging: Part 1. Curr Probl Diagn Radiol 2014;43(6):331-346.

List of figures

Figure number	Title
1	US HC plane
2	MRI HC plane

3	US BPD plane
4	MRI BPD plane
5	US AC plane
6	MRI AC plane
7	US FL plane
8	MRI FL plane
9	BA plot: Intraobserver BPD differences
10	BA plot: Interobserver BPD differences
11	BA plot: Intraobserver HC differences
12	BA plot: Interobserver HC differences
13	BA plot: Intraobserver AC differences
14	BA plot: Interobserver AC differences
15	BA plot: Intraobserver FL differences
16	BA plot: Interobserver FL differences
17	BA plot: Intraobserver EFW differences
18	BA plot: Interobserver EFW differences

Appendix: Reporting Checklist

GRRAS checklist for reporting of studies of reliability and agreement

Version based on Table I in: Kottner J, Audigé L, Brorson S, Donner A, Gajeweski BJ, Hróbjartsson A, Robersts C, Shoukri M, Streiner DL. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106

Section	Item #	Checklist item	Reported on page #
Title/Abstract	1	Identify in title or abstract that interrater/intrarater reliability or agreement was investigated.	
Introduction	2	Name and describe the diagnostic or measurement device of interest explicitly.	
	3	Specify the subject population of interest.	
	4	Specify the rater population of interest (if applicable).	
	5	Describe what is already known about reliability and agreement and provide a rationale for the study (if applicable).	
Methods	6	Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations.	
	7	Describe the sampling method.	
	8	Describe the measurement/rating process (e.g. time interval between repeated measurements, availability of clinical information, blinding).	
	9	State whether measurements/ratings were conducted independently.	
	10	Describe the statistical analysis.	
Results	11	State the actual number of raters and subjects/objects which were included and the number of replicate observations which were conducted.	
	12	Describe the sample characteristics of raters and subjects (e.g. training, experience).	
	13	Report estimates of reliability and agreement including measures of statistical uncertainty.	
Discussion	14	Discuss the practical relevance of results.	
Auxiliary material	15	Provide detailed results if possible (e.g. online).	